CC Investigation 5: Variability

**Mathematical Goals**

- Understand that information can be gained about a population by examining statistics of a representative sample of the population, where random sampling tends to produce representative samples.
- Draw inferences about a population based on data from a random sample.
- Generate or simulate multiple samples of the same size to gauge the variation in estimates or predictions.
- Informally assess the degree of visual overlap of two data distributions with similar variabilities and express the difference between centers of the distributions as a multiple of a measure of variability.

**Teaching Notes**

In this investigation, students will build on their previous work with single data distributions to compare two data distributions and answer questions about differences the distributions suggest about the populations. Students will begin informal work with random sampling to generate data sets and explore the importance of representative samples for drawing inferences.

**Problem 5.1**

*Before Problem 5.1, present the following scenario to students:*

*Plains Middle School is considering the following locations for a Grade 7 field trip: science museum, state park, or ballet company. The principal wants to survey a sample of students to find which location Grade 7 students would prefer. Ask:*

- **Should the principal select the members of the science club for the sample?** (No; not everyone in the population has an equal chance of being selected, and their opinions might be biased in favor of the science museum.)
- **Should the principal survey every tenth student to walk into the middle school on a given morning?** (No, because that sample would include students outside the population of seventh graders.)
- **Should the principal ask a group of randomly-selected Grade 7 girls?** (No, not everyone in the population has an equal chance of being selected.)
- **How can the principal get a random sample of Grade 7 students?** (Answers will vary, leading to a discussion of the principles of how to generate randomness in samples.)

*During Problem 5.1 B, ask: How do you find the mean of a set of data?* (Divide the sum of the data values by the number of values in the set.)
During Problem 5.1 C, ask:
• **How do you find the probability that any one filling will be randomly chosen?** (Divide 1 by the total number of choices, 6: \( \frac{1}{6} \))
• **What are the possible outcomes of tossing a number cube?** (1, 2, 3, 4, 5, and 6)
• **Why is an experiment with a number cube a good test of your conclusion in Part 1?** (A number cube also has 6 equally-likely outcomes.)

**Problem 5.2**

Before Problem 5.2, review with students how to read a box plot. Ask:
• **What was the least amount Tim earned?** ($35)
• **What was the greatest amount Tim earned?** ($115)
• **How do you find the interquartile range for data shown in a box plot?** (Subtract the lower quartile, indicated by the left side of the box, from the upper quartile, indicated by the right side of the box.)

During Problem 5.2 B, ask: **What do the range and interquartile range tell you about how data varies?** (The range sizes tell you how spread out the data are.)

After Problem 5.2 C, ask: **What does a symmetrical box plot tell you about how data vary?** (Data in the first and fourth quarters are distributed over the same range and data in the second and third quarters are distributed over the same range.)

During Problem 5.2 F, ask: **What does a cluster of data tell about a data set?** (More data is distributed over one small interval than over other intervals.)

**Problem 5.3**

During Problem 5.3 A, ask: **How are the distributions the same and how are they different?** (The interquartile ranges are the same and the medians are different.)

Before Problem 5.3 B, review with students how to read a dot plot. Ask:
• **What do the dots above the number 40 in the Bountiful Bistro plot represent?** (Five diners spent 40 minutes at dinner at Bountiful Bistro.)
• **How can you tell on each plot how long most of the diners spent at dinner?** (Find the number with the most dots above it.)
• **How can you find the median value?** (Count the total number of dots, then count dots beginning at the lowest value until you reach the middle value.)

During Problem 5.3 B Part 3, ask: **How does the size of the interquartile range translate visually onto a dot plot?** (The greater the interquartile range, the more spread out the data will appear.)
Summarize

To summarize the lesson, ask:

• **What is the best way to produce a sample that will support valid inferences about the population?** (a random sample in which every member of the population has an equal chance of being included)

• **How does the sample size affect how accurate predictions about the population sampled will be?** (The larger the sample size, the more reliable the results.)

• **What can you compare about two data sets just by looking at their distributions on a box or dot plot?** (medians and relative sizes of mean absolute deviations)

Students in the CMP2 program will further study standards 7.SP.1, 7.SP.2, and 7.SP.3 in the Grade 8 Unit *Samples and Populations.*

Assignment Guide for Investigation 5

Problem 5.1, Exercises 1–6, 12–14

Problem 5.2, Exercises 7–11, 15–18

Problem 5.3, Exercises 19–29

Answers to Investigation 5

**Problem 5.1**

A. 1. No; by selecting only women in business attire, members of the population who are male or not wearing business attire have no chance of being selected.

2. No; selecting diners only at lunch excludes breakfast or dinner diners who also are part of the population.

3. Survey every tenth diner all day. Each member of the population would have an equal chance of being selected.

B. 1. Casual Café: 38.4; Bountiful Bistro: 21.75; The mean for Casual Café is much greater, so that restaurant attracts older diners on average.

2. Casual Café: Because younger, college-age, diners are attracted to the restaurant, advertise in local college newspapers or on-line and by email;  
Bountiful Bistro: Since the restaurant attracts older diners, advertise in the newspaper or local magazines.

C. 1. 1 or 2 turkey sandwiches; In 6 orders,  
I would expect $\frac{1}{6}$, or 1, turkey sandwich,  
and in 12 orders I would expect $\frac{2}{12}$, or 2, turkey sandwiches.

2. Check students’ work.

3. $50 \times \frac{1}{6} = 8\frac{1}{3}$; about 8 turkey sandwiches;  
Check students’ work.

4. Check students’ work.

5. Check students’ answers; Student outcomes generally should be closer to theoretical probabilities with more trials.

**Problem 5.2**

A. Tim: range = 115 – 35 = $80$,  
interquartile range = 105 – 80 = $25$;  
Dan: range = 115 – 45 = $70$,  
interquartile range = 100 – 60 = $40$

B. The ranges show that Tim’s earnings, at their extremes, vary more widely than Dan’s. But the interquartile ranges show that Tim’s earnings are more clustered around the median value, which is higher than Dan’s.
C. Yes; the plot for Dan’s earnings is symmetric.

D. Dan’s earnings are more widely distributed, while Tim’s earnings are more clustered.

E. The plot for Tim’s earnings shows clustering about the median value, especially from 80 to 90.

F. Tim more consistently earns about the same amount in tips, while Dan’s earnings are much less consistent, though they do not vary as greatly in the extremes as Tim’s earnings.

G. Tim; the lower quartile of his tips is equal to the median of Dan’s tips, so the overall earnings should be greater.

**Problem 5.3**

A. 1. The values for Bountiful Bistro are generally less than the values for Casual Café, so the cost of dinner at Bountiful Bistro generally is less than the cost of dinner at Casual Café.

2. a. Bountiful Bistro: median = $15, interquartile range = 20 – 10 = $10; Casual Café: median = $25, interquartile range = 30 – 20 = $10

   b. 25 – 15 = $10

3. Yes, the median price at Casual Café is greater than the median price at Bountiful Bistro.

B. 1. Diners at Bountiful Bistro generally spend more time at dinner than diners at Casual Café.

2. Bountiful Bistro: median = 50 min; Casual Café: median = 20 min; 50 – 20 = 30 min

3. Bountiful Bistro; the data are more spread out for Bountiful Bistro, with a range of 70 min, than the data for Casual Café, with a range of 40 min.

**Exercises**

1. No; your friends are not representative of the population.

2. Yes; every 20th student is representative of the population.

3. Yes; students at different grades is representative of the population.

4. No; the soccer team is not representative of the population.

5. Sample: Randomly select names by choosing every fifth name from a list of names of students in your class.

6. about 4 minutes; Since the songs are randomly selected, each sample should yield close to the same results. A song length of about 4 minutes is between the means of the first two samples.

7. a. Casual Café: $16; Bountiful Bistro: $12.50

   b. Casual Café: $22.50; Bountiful Bistro: $20

   c. Casual Café: $7.50; Bountiful Bistro: $12.50

8. Customers at Casual Café generally spent more than customers at Bountiful Bistro.

9. At Casual Café, the difference between the least and greatest amounts is greater than at Bountiful Bistro, but the data are more clustered around the median.

10. The data for Casual Café are symmetric, so the amounts spent are evenly distributed. The data for Bountiful Bistro are not symmetric, so those amounts are less evenly distributed.

11. The data for Casual Café show clustering about the median, meaning more customers spend close to the median than customers do at Bountiful Bistro. The data for Bountiful Bistro show a cluster between $5.00 and $7.50; 25% of the purchases at this restaurant are in this range.

12. 17 cats

13. 32 dogs

14. 20 cats; 20 is closer to the mean for cats than the mean for dogs.

15. Casual Café is open 6 days each week and Bountiful Bistro is open 7 days each week. In the graph for Casual Café, there are 0 customers for Monday.

16. B

17. Both restaurants show peaks of numbers of customers on Saturdays.

18. The data for Casual Café are not symmetrically distributed, while the data for Bountiful Bistro are symmetric.

19. Sets A and C

20. Sets A and B

21. The review was effective because the students taking the review class scored higher than those who did not take the class. The box plots are the same size and shape, with the Review Class shifted right and showing higher scores.
22. 10%

23. The difference in medians is almost twice the mean absolute deviation, which shows there is substantial difference between the data sets.

24. Player A: mean = 26 points; Player B: mean = 18 points; difference = 8 points

25. Player A: MAD = 2.4; Player B: MAD = 2.2

26. about three-and-a-half times as great

27. No, the data values show no overlap.

28. Yes, the mean absolute deviation would be more than twice the difference in means, so there should be a lot of overlap.

29. No; the difference in means is so much greater than the mean absolute deviations that there shouldn’t be much overlap.
1. Rachel and Dwayne are running for Grade 7 President. Rachel is active in the drama club, and Dwayne plays on the school basketball team. Tell whether each sample would support a valid prediction about the outcome. Explain your reasoning.
   
a. a random sampling of members of the drama club
   
b. every member of the drama club and each player on the basketball team
   
c. a random sampling of Grade 7 girls
   
d. five randomly-selected students from each Grade 7 homeroom
   
e. every fifth student to walk into the middle school on a given morning

2. The box plots show the grades on history tests in two classes.

<table>
<thead>
<tr>
<th>Mr. Abrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 65 70 75 80 85 90 95 100</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mr. Philips</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 65 70 75 80 85 90 95 100</td>
</tr>
</tbody>
</table>

   a. What comparisons can you draw from the plots about the grades received by the two classes?
   
b. What is the range and interquartile range of the data displayed in each box plot?
   
c. What is the difference in the median values for the data sets?
   
d. Use the ranges, interquartile ranges, and medians to compare how the grades between the classes vary.
Skill: Representative Samples

Tell whether the sampling method will result in a representative sample. Explain your reasoning.

1. Nicole is planning the games for a carnival for the first, second, and third graders at Bay Elementary School. To find out which games students would like to have, she asks 25 first graders during their lunch.

2. Mr. Williams is deciding what books to offer for the next book club for his seventh- and eighth-grade English classes. He randomly asks ten students from each of his classes to make their choices from three different books.

3. A city is surveying its residents to find out if an open space should be developed into a park or an office building. The city sends surveys to 100 randomly-selected residents of the city.

4. Alyssa is doing research for a report about the after-school activities of students at her school. She interviews every fifth student entering the gym after school.

Skill: Reading Box Plots

Compare the medians and interquartile ranges of each pair of plots, and explain what conclusions you can make about the data sets based on those measures.
1. Brahim and Miguel are conducting a survey. They ask the question, “Which sport is your favorite to watch: soccer, basketball, or volleyball?”
   a. Brahim wants to ask 25 of his classmates as they leave a basketball game. Will his results be reliable? Explain.
   
   b. Miguel wants to ask 1 randomly-selected student from each of 5 gym classes. Will his results be reliable? Explain.
   
   c. Describe a sampling method Brahim and Miguel could use to have the best chance of producing a representative sample for their survey.

2. Riders on a subway system can get on one of four different subway lines. Rita works for the subway system and wants an accurate prediction of how many of the 10,000 riders who use the station each day take each line.
   a. Rita watches where 10 riders go, and sees that 2 take the green line. Based on Rita’s sample, how many riders would you predict take the green line each day? Explain your reasoning.
   
   b. Rita watches the next 10 riders, and sees that 6 of them take the green line. Based on this sample, how many riders would you predict take the green line each day? Explain your reasoning.
   
   c. Rita takes a third sample, of 200 riders, and sees that 70 of them take the green line. Based on this sample, how many riders would you predict take the green line each day? Explain your reasoning.
   
   d. Were your estimates different? If so, explain why they were different. Tell which estimate you think is most accurate and explain your reasoning.
3. Sarah and DeShawn work part-time at the bowling alley. The box plots show the number of hours they have worked each week this year.

a. Find the ranges and interquartile ranges of the hours Sarah and DeShawn have worked, and use them to compare how their hours varied.

b. Compare how the amounts of hours Sarah and DeShawn worked are distributed.

c. Does either box plot show clustering or symmetry of data? If so, what does that show about the numbers of hours worked?

4. The dot plots show the numbers of texts, to the nearest 10, some students sent one week in the summer and one week during the school year.

a. What comparisons can you draw from looking at the plots about the numbers of texts sent during each time of year?

b. What is the difference in the median value for each set of data?

c. For which set of data would you expect a greater interquartile range? Explain your answer.
You can collect data from a random sample of a given population and use that data to make inferences about the population as a whole. Inferences will be valid only if the sample is representative of the population.

A sample is **representative** if every member of the population has an equal chance of being included in the sample. Random sampling is the best way to produce a representative sample that will support valid inferences.

**Problem 5.1**

A. The owners of the Casual Café and the Bountiful Bistro want to know more about the types of customers that dine at their restaurants. They each conduct a survey to find their customers’ ages and the price they would expect to pay for an entrée.

1. Suppose the owners took their samples by surveying the first fifteen women dressed in business attire. Do you think this sample is representative of the population? Explain.

2. Suppose the owners took their samples by surveying every fifth customer at lunch. Do you think this sample is representative of the population? Explain.

3. Describe a survey method that would give the restaurant owners a representative sample of the population. Explain how you decided on your method.

B. The table shows age data the owners gathered from a representative sample at each restaurant.

<table>
<thead>
<tr>
<th>Casual Café</th>
<th>34</th>
<th>41</th>
<th>45</th>
<th>67</th>
<th>23</th>
<th>19</th>
<th>45</th>
<th>34</th>
<th>32</th>
<th>35</th>
<th>34</th>
<th>56</th>
<th>63</th>
<th>23</th>
<th>25</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bountiful Bistro</td>
<td>29</td>
<td>17</td>
<td>23</td>
<td>18</td>
<td>14</td>
<td>28</td>
<td>21</td>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. What is the mean age for each restaurant’s customers? What do the mean ages tell you about the customers that each restaurant attracts?

2. The owners are deciding how to advertise their restaurants. They want to advertise to the group of customers that they expect will want to dine with them. Give some recommendations to each owner about how and where they should advertise.
C. Diners at Casual Café can make their own sandwiches starting with 1 of 6 fillings.

1. If the sandwiches are randomly chosen, how many turkey sandwiches do you expect there will be in the next 10 sandwich orders? Explain how you found your answer.

2. Do an experiment to test your conclusion. Toss a number cube 10 times and record the outcomes in a table. Did the number of times you tossed a 3 match your prediction for the number of turkey sandwiches ordered? Explain why or why not.

3. How many turkey sandwiches would you expect out of the next 50 random sandwich orders? Do another experiment to test this conclusion. Toss a number cube 50 times and record the outcomes in a table.

4. Repeat the experiment for another 50 tosses. Record the outcomes in a separate table.

5. Were the experiments’ outcomes closer to your predictions for 10 orders or for 50 orders? Explain why that might be so.
You can use measures of variability, measures of center, and shape to compare the data displayed in two related graphs.

Problem 5.2

Tim is on the wait staff at the Casual Café, and Dan is on the wait staff at Bountiful Bistro. The box plots below display the amounts they earned in tips on weekends during the past six months.

**A.** What is the range and interquartile range of the data displayed in each box plot?

**B.** Use the ranges and interquartile ranges of Tim’s and Dan’s tips. Compare how their tips vary.

**C.** Are either of the box plots symmetric?

**D.** Compare how the amounts of Tim’s and Dan’s tips are distributed.

**E.** Which, if any, of the box plots shows clusters of data?

**F.** Use the evidence of clusters or no clusters to compare Tim’s and Dan’s tips.

**G.** Overall, who do you think earns more tip money? Explain.
You can use measures of variability, such as interquartile range and mean absolute deviation, to make sense of data sets, both numerically and visually.

**Problem 5.3**

A. The box plot compares the dinner ticket amounts for the two restaurants.

![Box Plot]

1. Compare the distributions of the data shown in the box plot. What conclusions can you draw about the cost of dinner?

2. a. Find the median value and the interquartile range for each restaurant.
   
   b. What is the difference in the medians?

3. Do the results you found support the conclusions you made about the data? Explain why or why not.

B. The dot plots show the lengths of time, to the nearest 10 minutes, some diners spent at dinner at each restaurant.

![Dot Plots]

1. What comparisons can you draw from looking at the plots about the time diners spend having dinner at the restaurants?

2. What is the difference in the median value for each set of data?

3. For which set of data would you expect a greater interquartile range? Explain your answer.
Exercises

For Exercises 1–4, tell whether the sample is representative of the population.

1. You want to know what type of music students at your school like best. You ask a group of your friends which music they like best.

2. You want to know which type of food students at your school like best. You ask every 20th student in your school yearbook.

3. You want to know how many hours students at your school spend on the computer each day. You ask students from different grades as they leave school.

4. You want to know how many hours students at your school exercise each week. You ask the members of the soccer team how often they exercise each week.

5. Suppose you are taking a poll of students in your grade to see whom they are going to select in the election for president of your class. Describe one way you could find a sample that is representative of the population.

6. A student is trying to determine the average length of a song in her large music library. She randomly selects 20 songs and finds that the mean length is 4 minutes 9 seconds. Then, she randomly selects another 20 songs and finds that the mean length is 3 minutes 52 seconds. What would you expect the mean length of a third set of 20 songs would be? Why?
For Exercises 7–11, use the data displayed in the box plots below.

Amount Spent by Each Customer (in dollars),
Casual Café

Amount Spent by Each Customer (in dollars),
Bountiful Bistro

7. Find the following for each set of data.
   a. median
   b. range
   c. interquartile range

8. Use the medians of the data to compare the amounts spent by customers at each restaurant.

9. Use the ranges and interquartile ranges of the data to compare how the amounts spent by customers at each restaurant vary.

10. Use the symmetry or lack of symmetry in each box plot. Compare how the amounts spent by customers at each restaurant are distributed.

11. Use the evidence of clusters or no clusters to compare the amounts spent by customers at each restaurant.

For Exercises 12–14, use the information about the numbers of cats and dogs that were adopted at a local shelter each month last year.

<table>
<thead>
<tr>
<th>Cats</th>
<th>12</th>
<th>13</th>
<th>16</th>
<th>18</th>
<th>21</th>
<th>15</th>
<th>14</th>
<th>13</th>
<th>15</th>
<th>22</th>
<th>19</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dogs</td>
<td>25</td>
<td>30</td>
<td>38</td>
<td>29</td>
<td>27</td>
<td>40</td>
<td>33</td>
<td>26</td>
<td>32</td>
<td>34</td>
<td>41</td>
<td>29</td>
</tr>
</tbody>
</table>

12. What is the mean number of cats adopted each month?

13. What is the mean number of dogs adopted each month?

14. A worker knows that either 20 cats or 20 dogs were adopted one month recently. Based on your answers to Exercises 12 and 13, do you think it was 20 cats or 20 dogs? Explain.
For Exercises 15–18, use the data displayed in the bar graphs below.

15. Compare the numbers of days the restaurants are open. Explain how the graphs show this.

16. **Multiple Choice** Which statement is true about the mean numbers of customers during days that each restaurant is open?
   
   A. The mean number of daily customers at Bountiful Bistro is about 20 more than at the Casual Café.
   
   B. The mean number of daily customers at Bountiful Bistro is about 10 more than at the Casual Café.
   
   C. The mean number of daily customers at the restaurants is about the same.
   
   D. The mean number of daily customers at Bountiful Bistro is about 10 less than at the Casual Café.

17. Use any peaks in the data to compare the numbers of customers at the restaurants.

18. Use any symmetry or lack of symmetry to compare the distribution of data for the restaurants.

For Exercises 19–20, determine which two sets of data will overlap more—Sets A and B or Sets A and C.

19. Set A has a mean of 12 and a mean absolute deviation of 5.1.  
   Set B has a mean of 23 and a mean absolute deviation of 4.9.  
   Set C has a mean of 10 and a mean absolute deviation of 4.8.

20. Set A has a mean of 104 and a mean absolute deviation of 19.6.  
   Set B has a mean of 84 and a mean absolute deviation of 25.  
   Set C has a mean of 180 and a mean absolute deviation of 20.
For Exercises 21–23, use the information given about class test scores shown in this box plot.

21. What conclusions can you draw from looking at the plot about how effective the math exam review class was?

22. What is the difference in the medians between the sets of data?

23. The mean absolute deviation for both groups of students is 6.2. Compare that value to the difference in medians. What does that tell you about the data?

For Exercises 24–28, use the information given about the points that two basketball players scored in each of the games they played this year.

<table>
<thead>
<tr>
<th>Player A</th>
<th>30 26 21 28 24 28 25 26 30 22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Player B</td>
<td>16 18 15 18 22 14 16 23 18 20</td>
</tr>
</tbody>
</table>

24. Find the mean number of points scored for each player. Find the difference in the means.

25. Find and compare the mean absolute deviation for each player.

26. How many times greater is the difference in the means than the mean absolute deviation for each player?

27. Would you expect there to be a lot of overlap in dot plots of the data? Why or why not?

28. Suppose the mean number of points scored for Player B were 25 points, and the variability stays the same. Would you expect there to be a lot of overlap in a dot plot of the data? Why or why not?

29. Your mean quiz score is 15 points higher than your friend’s mean quiz score, which is 3 times the mean absolute deviation of both of your scores. Do you think there will be a lot of overlap if you make a double histogram of the data? Explain.